Annotation sémantique de documents pour la révision des règles métiers

Abdoulaye Guissé, François Lévy, Adeline Nazarenko, Sylvie Szulman

Annotation sémantique pour l'indexation de règles métiers ⋆

A. Guissé, F. Lévy, A. Nazarenko, S. Szulman

LIPN UMR 7030 (Université Paris 13 & CNRS)
99, av. J.B. Clément, 93430, Villetaneuse
prénom.nom@lipn.univ-paris13.fr

Résumé : Les systèmes décisionnels reposent généralement sur un ensemble de règles métiers dont la formalisation nécessite souvent de revenir aux documents sources pour des raisons de justification ou de maintenance. Cela pose un problème complexe d'annotation sémantique puisqu'il faut articuler des textes réglementaires, des règles métiers qui en sont dérivées de manière plus ou moins directe et une ontologie du domaine qui décrit le vocabulaire conceptuel dans lequel les règles métiers s'expriment. Nous proposons une structure riche d'index qui permet de passer des concepts de l'ontologie et des règles aux textes et inversement.

Mots-clés : Annotation, indexation, règles métiers, ontologie.

1 Introduction

Les systèmes décisionnels reposent généralement sur un ensemble de règles métiers qui décrivent formellement les critères et les processus de décision. L'une des difficultés consiste à traduire les réglements qui n'existent souvent que sous la forme rédigée en un système cohérent et complet de règles formalisées. Les méthodes d'acquisition de connaissances à partir de textes et d'extraction d'information ne permettent pas de dériver automatiquement ces règles des textes réglementaires qui sont pourtant de précieuses sources d'information. L'édition des règles doit donc se faire au moins en partie manuellement.

Pour assister ce travail d'édition, il est important d'expliciter l'ontologie du domaine, ce qui détermine le vocabulaire conceptuel à utiliser. Il est également important de s'appuyer sur les documents sources qui contiennent les connaissances à formaliser sous la forme de règles. Une fois les règles construites, il est précieux de conserver le lien entre une règle et le passage de texte dont elle est dérivée pour pouvoir justifier les règles elles-mêmes ou les prises de décision. C'est également utile en termes de maintenance : si la documentation évolue il

⋆. Ce travail a été partiellement financé dans le cadre du projet Ontorule FP7 Collaborative project n°231875.

faut que la base de règles soit mise à jour en conséquence (suppression ou ajout de règles) ; inversement, si la base de règles est modifiée, il faut indiquer quels textes doivent être actualisés.

La gestion des règles métiers dans les systèmes de décision pose ainsi un problème complexe d'annotation sémantique : il faut articuler des textes réglementaires, des règles métiers qui en sont dérivées de manière plus ou moins directe et une ontologie du domaine qui décrit le vocabulaire conceptuel dans lequel les règles métiers s'expriment. C'est ce problème d'annotation sémantique que nous cherchons à résoudre. Nous proposons ici une structure riche d'index qui permet de passer des concepts de l'ontologie et des règles aux textes et inversement.

Après avoir situé notre travail par rapport aux travaux d'annotation sémantique existants et dans une problématique d'indexation (section 2), nous présentons notre modèle d'index en montrant en quoi il étend les approches classiques d'annotation et comment il est implémenté (section 3). L'approche proposée a été testée sur un problème particulier de gestion des points de fidélité par une compagnie aérienne. Les résultats de cette première expérimentation sont présentés dans la section 4.

2 De l'annotation à l'indexation

Le terme d'« annotation sémantique » désigne aussi bien l'activité consistant à apposer une « note » sur une partie de document ou de texte que la note qui en résulte. Il renvoie d'emblée à une multitude de pratiques depuis les remarques des relecteurs qui viennent commenter ou corriger les textes qu'ils lisent, jusqu'aux clés d'indexation apposées par les documentalistes, aux « tags » des usagers du web 2.0 ou aux propriétés linguistiques qui peuvent être explicitées pour faciliter le retravail de certains textes. De façon générale, l'annotation consiste en un apport d'informations de nature interprétative aux données brutes (Leech, 1997). L'annotation sémantique se traduit par la définition d'une sur-couche d'informations sémantiques qui viennent donner un sens aux textes. Ce sont des « meta-données », étant entendu que ces méta-données peuvent avoir une portée locale et être seulement relatives à des fragments de texte.

2.1 Outils d'annotation sémantique

Les outils qu'on appelle « d'annotation sémantique » se situent généralement dans le cadre du Web Sémantique. Ils servent à créer et gérer des annotations qui donnent une description formelle du contenu des ressources du Web. Ils reposent sur un modèle formel de connaissances, en général une ontologie, et exploitent de plus en plus les standards du Web Sémantique (principalement XML pour les documents, SKOS et OWL pour le modèle sémantique, RDF pour les annotations).

Il existe de nombreux outils d'annotation sémantique [1] qu'Amardeilh (2007)

1. Voir (Uren *et al.*, 2006) pour une revue de l'état de l'art.

distingue selon la nature des ressources documentaires annotées (texte, image, vidéo, etc.), le mode d'annotation (automatique, semi-automatique, ou manuel), l'ontologie de référence utilisée, etc. Nous nous intéressons ici aux seuls outils d'annotation de documents textuels.

Les outils les plus fréquemment utilisés pour l'annotation de documents textuels sont SMORE (Kalyanpur *et al.*, 2003), Annotea (Kahan *et al.*, 2001), Semtag (Dill *et al.*, 2003), KIM (Popov *et al.*, 2003), UIMA (IBM, 2006), ce dernier étant plus une plate-forme incorporant des modules d'annotation sémantique. Ces outils reposent eux-mêmes sur des outils d'extraction pour identifier les fragments de texte à annoter et leur associer une étiquette sémantique. Il s'agit d'ordinaire de systèmes d'extraction d'informations dans des ressources non structurées qui exploitent une analyse linguistique du texte comme GATE (Cunningham, 2002), Amilcar (Ciravegna & Wilks, 2003) ou même UIMA (IBM, 2006), en tant que plate-forme intégrant des modules d'extraction.

Du point de vue sémantique, le processus d'annotation se traduit souvent par un peuplement d'ontologie avec la détection de nouvelles instances de concepts ou de relations entre instances qui viennent enrichir l'ontologie (Popov *et al.*, 2003; Amardeilh *et al.*, 2005). Ce sont alors les entités nommées qui sont repérées dans les textes et annotées par les instances de concepts auxquelles elles renvoient, les entités nommées étant des unités textuelles dotées d'une autonomie référentielle.

2.2 Annoter pour indexer

Les usages de la représentation sémantique issue de l'annotation peuvent à grands traits être séparés en deux classes. Dans la première, l'annotation est utilisée pour sa valeur sémantique et sa source textuelle n'est plus nécessaire une fois que le travail d'analyse a été fait. Il s'agit alors d'extraire des connaissances, éventuellement d'alimenter des bases de données. Dans la seconde, l'annotation sert à accéder au texte qui lui a donné naissance. Nous parlons alors d'indexation et c'est ce type d'utilisation qui nous intéresse ici.

La représentation sémantique peut en effet être utilisée à des fins de recherche documentaire, le lien entre annotations et ressources textuelles facilitant l'interrogation de ces dernières. Dans une perspective d'acquisition de connaissances à partir de textes, les annotations sémantiques permettent de tracer le lien entre des documents et des ressources sémantiques construites à partir de ces derniers. L'indexation permet aussi de documenter les connaissances que représentent les annotations sémantiques, voire de les maintenir à jour quand les textes de référence évoluent. De manière générale, il s'agit d'utiliser la structure sémantique construite par les annotations pour identifier des éléments dans un document et naviguer des éléments sémantiques aux fragments de texte ou vice-versa.

Le terme d'« indexation » est utilisé ici pour désigner l'ensemble des annotations sémantiques mais vues comme un espace de navigation entre les textes et une structure sémantique. Nous parlons d'« index » pour désigner la structure complexe qui associe une structure sémantique à un texte *via* des annotations sémantiques particulières.

2.3 Structure d'un index

Plus formellement et de manière générale, nous définissons un index comme une structure composée de 3 sous-structures.

Le *modèle de document* détermine quels fragments de texte sont des unités documentaires pouvant supporter un lien d'indexation. En principe n'importe quelle liste de caractères peut être annotée mais on se limite en général à des intervalles continus. On peut aussi typer les séquences de caractères en question pour distinguer des mots, des syntagmes, des phrases, des paragraphes, des sections de document, etc. La structure d'index dépend largement du modèle de document considéré. Dans le cas de la plateforme KIM (Popov *et al.*, 2003), par exemple, seules les entités nommées peuvent servir de support à l'indexation. Dans un index traditionnel comme celui de la base MedLine [2], l'unité documentaire généralement considérée est le document pris dans son ensemble.

Le *modèle sémantique*, quant à lui, indique quelles unités sémantiques peuvent être associées aux unités documentaires et quelles relations ces unités sémantiques entretiennent entre elles. Ce modèle sémantique peut être un simple thesaurus (les documents de MedLine sont associés à des descripteurs extraits du thesaurus MESH [3]) mais, dans le cadre du Web Sémantique, il s'agit généralement d'un modèle ontologique. Il est souvent utilisé de manière partielle, cependant, comme dans KIM où seules des instances peuvent être utilisées pour annoter [4] alors qu'on peut généraliser l'approche et donner tout élément de l'ontologie (concept, rôle, instance de concept ou de rôle) comme cible du lien d'indexation.

Le *modèle de correspondance* associe des unités documentaires (ud), à des unités sémantiques (us). Dans le cas le plus simple, un lien d'indexation se représente comme un couple (ud_i, us_i) mais ces liens peuvent porter des propriétés : on peut les typer pour marquer le rôle que le fragment de texte joue par rapport à l'unité sémantique (définition *vs.* exemple), indiquer l'usage pour lequel ils sont proposés (spécialisé *vs.* grand public), leur associer un poids de pertinence, etc. Dans le cas général, un lien d'indexation se représente donc comme un triplet (l_i, ud_i, us_i) où l_i est la liste des propriétés du lien associant ud_i à us_i.

3 Un modèle d'indexation

Les outils d'annotation souffrent généralement d'une double limitation qui contraignent la richesse du modèle d'indexation qu'ils peuvent supporter. Les types d'annotations possibles sont peu variés et le modèle sémantique retenu, l'ontologie, limite par lui-même le type des annotations qui peuvent être posées sur le texte. Dans notre cas, nous cherchons à associer divers types d'éléments ontologiques (instances, concepts, rôles) aux fragments textuels mais aussi des règles métiers qui ne sont pas toutes représentables dans une ontologie. Ceci nous invite à travailler sur un modèle sémantique étendu.

2. www.ncbi.nlm.nih.gov/pubmed
3. www.nlm.nih.gov/mesh
4. Elles sont créées à la volée lors de l'annotation pour venir peupler l'ontologie.

3.1 Modèle pour l'indexation des règles métier

Pour intégrer la documentation technique dans le système de gestion des règles métiers d'un outil d'aide à la décision, il faut s'appuyer sur une structure d'index assez riche.

Le document est représenté de manière classique comme une structure arborescente $T = \langle r, A_1^1, A_2^1, ...A_n^1 \rangle$ où r, la racine de l'arbre, représente le corpus complet qui s'analyse à la profondeur 1 en une séquence de sous-arbres A_i^1 correspondant à autant de sous-structures textuelles. Celles-ci s'analysent elles-mêmes récursivement en séquences ordonnées de structures plus élémentaires. Les noeuds de l'arbre correspondent donc aux structures suivantes : le corpus (la racine), le document, les différents niveaux de sections et sous-sections, les paragraphes, les phrases et les mots [5]. Une unité documentaire $ud(t)$ est bien formée si elle correspond à une séquence de noeuds de l'arbre relevant d'un même père. Autrement dit, $ud(t)$ est une unité documentaire ssi $ud(t) = r \vee (\exists k, l)(ud(t) = (A_i^k, A_{i+1}^k, ...A_{i+j}^k) \wedge A_l^{k-1} = (A_1^k, A_2^k, ...A_i^k, A_{i+1}^k, ...A_{i+j}^k, ...A_n^k))$ avec $j \geq 0$ et $n \geq i + j$. Concrètement, ce modèle documentaire autorise l'annotation d'une séquence de mots, d'une phrase, d'une séquence de paragraphes ou d'une section mais pas d'une liste de phrases qui commencerait au milieu d'un paragraphe et se poursuivrait sur le paragraphe suivant ou de deux mots disjoints.

Notre modèle sémantique présente une double originalité. Il tire profit de la diversité des éléments ontologiques (concepts ou instances de concepts le plus souvent) et il comporte, outre l'ontologie, une base de règles qui peuvent elles aussi être la cible d'un lien d'indexation. Certaines règles peuvent s'exprimer comme des restrictions de rôles dans l'ontologie mais elles sont néanmoins réifiées dans la base de règle, ce qui permet d'y faire référence. D'autres (les règles procédurales notamment) débordent le pouvoir de représentation d'une ontologie. L'ontologie et la base de règles constituent un modèle sémantique unifié. Etant donné l'ontologie $O = \langle C, R, I, RI \rangle$ composée d'un ensemble de concepts (C), rôles (R), instances (I) et relations entre instances (RI) ainsi que de la base de règles $BR = \{r_1, r_2, ...r_n\}$, toute unité sémantique us de $C \cup R \cup I \cup RI \cup BR$ peut être cible de liens d'indexation.

A ce stade, notre système repose sur un modèle de correspondance très simple qui se décrit comme un ensemble de couples associant une unité documentaire à une unité sémantique [6] $(C = \{(ud_1, us_1)(ud_2, us_2)), ...(ud_n, us_n)\})$.

3.2 Implémentation

Par souci d'uniformité et de compatibilité, nous utilisons pour les formats de représentation les standards du W3C pour le Web Sémantique.

Le texte est en XML. Il est découpé hiérarchiquement en document, sections, paragraphes, etc. La phrase est l'élément de plus bas niveau, l'élément mot restant implicite. Nous avons par exemple pour la phrase 11 d'un texte donné :

5. La notion de phrase ou de mot n'est pas définie ici sur des critères linguistiques mais formellement par algorithme de segmentation déterministe.

6. Noter que tous les ud_i et tous les us_j ne sont pas forcéments distincts entre eux.

```
<Sentence rdf:ID="11"><content>Each qualifying activity extends the
expiration date of all unexpired mileage credit in your account for
18 months from the date of the qualifying activity.</content>
</Sentence>
```

L'ontologie est représentée en OWL, Ontologie Web Language (Hendler *et al.*, 2004). Le concept QUALIFYING ACTIVITY qui est mentionné deux fois dans la phrase précédente est représenté par la classe OWL `Qualifying_activity`, sous-classe de `Activity` :

```
<owl:Class rdf:ID="Qualifying_activity">
  <rdfs:subClassOf> <owl:Class rdf:about="#Activity"/>
  </rdfs:subClassOf>
</owl:Class>
```

Les règles métiers sont définies en RIF (Rule Interchange Format, RIF06), standard du Web pour la définition de règles. L'exemple ci-dessous utilise la "RIF-Core Presentation Syntax" (traductible en xml par sérialisation) pour définir la règle R6 décrite par la phrase précédente. L'opérateur `:-` est l'opérateur d'implication des règles de production.

```
Prefix(func <http://www.w3.org/2007/rif-builtin-function#>)
Prefix(terminae http://lipn.univ-paris13.fr/terminae#)
forall(R6) ?x ?y ?z ?date
  ?date [func:numeric-add -> "18"] :-
  AND(
    ?x [rdf:type -> terminae:Qualifying_activity
      terminae:hasAccount -> ?y]
    ?y [rdf:type terminae:Account]
    ?z [rdf:type -> terminae:Mileage_credit
      terminae:unexpired -> "yes"
      terminae:isContentOf -> ?y
      terminae:expiration_date -> ?date]
    )
```

L'index est décrit sous forme de triplets, en RDF, Resource Description Framework (Ora & Swick, 1999). Dans l'exemple suivant, nous définissons deux correspondances. La première fait le lien entre la classe OWL `Qualifying_activity` et le fragment «qualifying activity» situé entre les positions 5 et 24 de la phrase 11. La seconde fait le lien entre la règle R6 et la phrase 11 dans son ensemble.

```
<!--Correspondance fragment de texte et concept d'ontologie-->
<Qualifying_activity rdf:ID="qualif_act_2">
  <string>qualifying activity</string>
  <start_offset>5</start_offset>
  <end_offset>11</end_offset>
  <hasSentence rdf:resource=
             "http://lipn.univ-paris13.fr/terminae-data#11"/>
</Qualifying_activity>
```

```
<!-- Correspondance fragment de texte et règle métier -->
<R6 rdf:ID="Regularity_rule_1">
  <hasSentence rdf:resource=
             "http://lipn.univ-paris13.fr/terminae-data#11"/>
</R6>
```

Cet index permet de naviguer d'une ressource à l'autre mais il peut être aussi chargé dans un moteur de recherche sémantique pour être interrogé *via* des requêtes SPARQL (Prud'hommeaux & Seaborne, 2006), un langage de requête pour le noyau commun RDF de nos formalismes sémantiques. Un tel dispositif permet par exemple de calculer, pour une règle métier donnée, la liste des concepts de l'ontologie et le texte auxquels elle est associée.

4 Expérience

Nous avons indexé un premier corpus pour montrer l'intérêt de cette structure de navigation.

4.1 Présentation du corpus

Le corpus choisi concerne un système de points de fidélité ou *avantages* destiné aux voyageurs utilisant régulièrement les services de la compagnie American Airline. Il est représentatif d'une large classe de textes décrivant les règles métier, tout en restant accessible au public. Il décrit les droits et obligations des parties, soit un peu plus de 5300 mots répartis en 256 paragraphes.

Chaque type d'avantage est décrit avec les conditions d'obtention des points, leurs conditions d'utilisation, leur mode de calcul, leur période de validité. La plupart des sections consistent en une liste de sujets traités indépendamment, chacun dans un paragraphe (1 à 9 lignes). L'annotation pertinente pour une règle embrasse donc au plus un paragraphe, le plus souvent une ou deux phrases.

4.2 Exemple détaillé

Considérons par exemple un fragment du premier paragraphe du texte :

> AAdvantage members **must** have mileage earning or redeeming activity once every 18 months **in order to** retain their miles. **Each** qualifying activity extends the expiration date of all unexpired mileage credit in your account for 18 months from the date of the qualifying activity. Qualifying activity **is defined as** redeeming any AAdvantage award or accruing mileage credit on any eligible American, American Eagle, AmericanConnection or AAdvantage airline participant **as well as** accruing mileage credit with participating hotels, car rental companies, credit cards, telecommunication providers, and other service providers offering AAdvantage mileage credit.

Les principaux termes renvoyant à l'ontologie du domaine sont soulignés. On y trouve ainsi mentionné le concept de BONUS DISPONIBLE (*credit*, *miles*, *mileage* ou *mileage credit*) qui a une DATE LIMITE (*expiration date*). Il subsume le concept de BONUS INUTILISÉ (*unused mileage credit*) dont une sous-classe est BONUS HORS DÉLAI (*expired milage credit*). Le BONUS GAGNÉ (*earned mileage* ou *accrued mileage*) est lié à un ACHAT (de *ticket* ou *transaction*) qui doit être CONVENABLE (*eligible ticket.* et *eligible participant*) même si ce lien n'est pas désigné par un nom ou un verbe spécifique. On relève aussi des entités nommées (*American Airline*, *American Eagle*, *AmericanConnection*) qui désignent des compagnies aériennes et qui peuvent se modéliser comme instances d'un concept COMPAGNIE AÉRIENNE.

L'analyse du texte fait également apparaître des éléments grammaticaux qui peuvent servir de marqueurs de règles : ils sont en gras dans notre exemple. Si *must ... in order to* et *is defined as* semblent assez fiables, *as well as* ne porte pas de sémantique par lui-même. *Each* peut marquer une régularité, mais reste un indice incertain. Certains concepts pourraient aussi jouer le rôle de marqueur de règle, par exemple dans cette phrase DATE LIMITE (*expiration date*).

Cette même analyse montre qu'on peut grossièrement catégoriser les règles sur la base de deux attributs : la valeur de la règle (obligation, recommandation, permission, necessité, affirmation de compétence, définition, règle de calcul, etc.) et sa portée (universelle, éventuelle, sous condition, par exception, etc.).

L'une des difficultés de la modélisation des règles métiers à partir de ce type de texte destiné aux clients de l'entreprise est qu'il faut renverser la perspective et exprimer les règles du point de vue de l'entreprise (le système de décision guide le comportement de l'entreprise et pas celui de ses clients, même si il doit respecter les règles déclarées à ceux-ci). Ainsi la première phrase décrit un état de fait (si les membres n'ont pas d'activité, leur crédit est perdu) et se rattache à la catégorie NECESSITÉ SOUS CONDITION alors que du point de vue des membres, elle s'interpréterait plutôt comme une obligation. La seconde phrase décrit au contraire une OBLIGATION SOUS CONDITION, puisque l'entreprise doit ajuster la date limite dès que le membre a une activité qualifiante. La dernière phrase étant plutôt une DÉFINITION, elle doit s'exprimer dans l'ontologie : QUALIFYING ACTIVITY serait ainsi un concept plus spécifique que ACTIVITY en lien avec des activités à définir (*redeeming an award* et *accruing mileage credit*).

4.3 Espace de navigation

Pour manipuler l'index au niveau sémantique et exploiter la richesse de sa structure, nous avons conçu une interface de navigation dans cette structure. Cet espace de navigation se décompose en trois zones principales comme le montre la figure 1 (nous reviendrons plus loin sur la zone 4). Le texte est affiché dans une fenêtre centrale (zone 2), phrase par phrase dans cette version préliminaire. Le modèle sémantique est ici présenté en deux parties : la zone 1 à gauche donne accès à l'ontologie et les règles sont visibles dans la zone 3 à droite. Ce modèle est cependant unifié dans la mesure où les règles sont liés aux éléments

de l'ontologie. Cela n'est pas visible sur la figure mais les occurrences du concept sélectionné dans l'ontologie sont automatiquement marquées en couleur dans le texte de même que l'unité documentaire associée à une règle (la phrase dans cette première version).

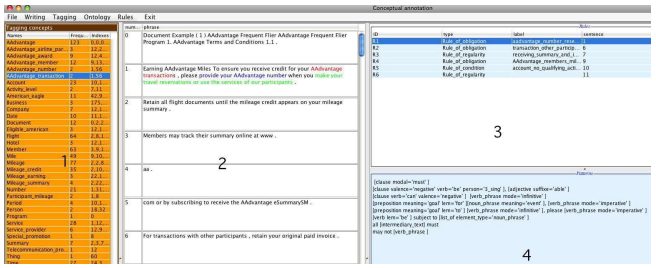

FIGURE 1 – Visualisation de l'espace de navigation

4.4 Construction de l'index

La construction de l'index suppose la reconnaissance d'une valeur sémantique dans un élément du modèle textuel. A terme, il faudra aussi reconnaître la force du lien lui-même. Nous distinguons ici les annotations ontologiques et les annotations de type règles.

Les méthodes de construction d'ontologies à partir de textes comme Terminae (Aussenac-Gilles *et al.*, 2008) ont l'avantage de conserver la trace des éléments textuels à partir desquels les concepts et les relations conceptuelles ont été construits. Se trouvent ainsi associés aux concepts les termes sous la forme desquels ils s'expriment dans les textes. On peut aussi avoir des marqueurs ou des patrons associés aux relations conceptuelles (Jacques & Aussenac-Gilles, 2006). Ce sont ces connaissances linguistiques associées à l'ontologie qui sont exploitées pour l'annotation des textes et la construction des liens d'indexation texte-ontologie.

L'annotation des règles métiers est plus complexe. L'analyse de corpus qui précède montre que cette annotation ne peut être faite automatiquement. Il faut donc prévoir un éditeur de règles, ce qui amène à penser l'index non pas uniquement pour la consultation des ressources (textes, ontologie, base de règle) mais aussi pour la construction des règles. Les analystes qui écrivent les règles métiers s'appuient en effet sur les sources textuelles. Il s'agit à la fois de localiser les zones de texte porteuses d'informations réglementaires et de les « traduire » en règles. Cela peut se faire par le repérage de marqueurs linguistiques et de schémas de phrases. Le vocabulaire déontique est un indice important, car une règle est en général une obligation ou une interdiction. Celle-ci peut se manifester dans le verbe (*devoir, être obligé de, être interdit*), une nominalisation (*obligation*) ou une qualification (*obligatoire*). On peut même reconnaître des structures caractéristiques de certaines catégories de règles même si ces structures restent souvent

sous-spécifiées et ambigües. Par exemple, une phrase de type «[NP] must [VP] in order to [VP]» a des chances d'être « traduite » en règle d'obligation.

Il faut donc, pour la construction des annotations elles-mêmes, prendre en compte des connaissances linguistiques associées aux éléments du modèle sémantique : des termes associés aux concepts, des entités nommées associées aux instances mais aussi des marqueurs et patrons associés aux règles et dans le cas général des règles d'annotation permettant de repérer, délimiter et désambiguïser les éléments du texte à annoter [7]. Nous considérons que ces éléments linguistiques ne font pas partie du modèle sémantique tel que défini dans la section 3.1 : les patrons de règles ne font pas plus partie de la base de règles que les termes ou marqueurs de relations de l'ontologie à proprement parler. Pour l'instant le lien entre les unités sémantiques et leurs réalisations linguistiques (éléments lexicaux ou patrons plus complexes) est maintenu de manière *ad hoc* – et on voit apparaître, dans la figure 1, une zone 4 qui présente des patrons associés aux catégories de règles de la zone 3 –, mais un modèle permettant de formaliser cette articulation est à l'étude (Ma *et al.*, 2009).

5 Conclusion

Cet article propose un modèle d'annotation sémantique et une première expérience d'annotation qui montrent comment on peut articuler des textes réglementaires, la base de règles métiers qui en est issue et une ontologie qui fixe le vocabulaire conceptuel dans lequel les règles sont exprimées. La structure obtenue est un index, et une interface de navigation permet de passer du texte à l'ontologie, du texte aux règles, des règles à l'ontologie, etc. Cette interface repose sur des standards du W3C.

De manière classique, notre modèle d'index se décrit comme l'association d'un modèle documentaire, d'un modèle sémantique et d'un modèle de correspondance qui relie des unités documentaires à des unités sémantiques. C'est la richesse de ce modèle qui en fait l'originalité. Une grande variété d'unités documentaires peuvent être source des liens d'indexation (du mot à la séquence de sections) et une grande variété d'unités sémantiques peuvent en être la cible (des instances de concepts mais aussi des concepts, des rôles ou leurs instances voire des règles métiers puisque notre modèle sémantique combine une ontologie et une base de règles métiers).

Références

Amardeilh F. (2007). Web sémantique et informatique linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle. In *Thèse de doctorat, Univ. Paris X*, p. 223–253.

7. La mise au point de ce ces règles d'annotation (ou patrons d'extraction) est une question reconnue comme délicate dans le domaine de l'extraction d'information et elle déborde du cadre de ce travail. Pour l'instant, nous nous contentons de prendre en compte des patrons lexicaux simples.

Amardeilh F., Laublet P. & Minel J.-L. (2005). Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques. In *Actes des 16èmes journée francophones d'Ingénierie des Connaissances*, p. 25–36.

Aussenac-Gilles N., Despres S. & Szulman S. (2008). The terminae method and platform for ontology engineering from texts. In P. Buitelaar & P. Cimiano, Eds., *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*. IOS Press.

Ciravegna F. & Wilks Y. (2003). Designing adaptive information extraction for the semantic web in amilcare. In H. S. & S. S., Eds., *Annotation for the Semantic Web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*, p. 112–127. IOS Press, Springer-Verlag.

Cunningham H. (2002). Gate - a general architecture for text engineering. In *Computers and the Humanities, Volume 36*, p. 223–254.

Dill S., Eiron N., Gibson D., Gruhl D., Guha R., Jhingran A., Kanungo T., S.Rajagopalan, Tomkins A., J.A.Tomlin & Zien J. (2003). Semtag and seeker : Bootstrapping the semantic web via automated semantic annotation. In *WWW'03*, p. 178–186, Budapest,Hongrie : ACM Press.

Hendler J., Horrocks I. & al. (2004). Owl web ontology language reference. In *W3C Recommendation*.

IBM (2006). Unstructured information management architecture (uima), sdk user's guide and reference. In *http ://dl.alphaworks.ibm.com/technologies/uima/UIMA_SDK_Users_Guide_Reference.pdf*, p. 364.

Jacques M.-P. & Aussenac-Gilles N. (2006). Variabilité des performances des outils de tal et genre textuel. cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues (TAL)*, **47**(1), 11–32.

Kahan J., Koivunen M., Prud'hommeaux E. & Swick. R. (2001). Annotea : An open rdf. In *Proceedings of the 10th Infrastructure for Shared Web Annotations WS (WWW'01)*, p. 623–632, Hong-Kong : ACM Press.

Kalyanpur A., Hendler J., Parsia B. & Golbeck J. (2003). Smore - semantic markup, ontology, and rdf editor. In *http ://www.mindswap.org/papers/SMORE.pdf*.

Leech G. (1997). Introduction to corpus annotation. In R. Garside, G. Leech & A. McEnery, Eds., *Corpus annotation : Linguistic information from computer text corpora*. Longman 1 : 18.

Ma Y., Audibert L. & Nazarenko A. (2009). Ontologies étendues pour l'annotation sémantique. In F. L. Gandon, Ed., *Actes des 20es Journées Francophones d'Ingénierie des Connaissances (IC 2009)*, p. 205–216, Hammamet, Tunisie : PUG.

Ora L. & Swick R. (1999). Resource description framework (rdf) model and syntax specification. In *16èmes journée francophones d'Ingénierie des Connaissances*. W3C Recommendation.

Popov B., Kiryakov A., Manov D., Kirilov A., Ognyanoff D. & Goranov M. (2003). Towards semantic web information extraction. In *Proceedings of the Human Language Technologies Workshop (ISWC'03)*, p. 1–22, Sanibel, Floride.

Prud'hommeaux & Seaborne E. (2006). Sparql query language for rdf. In *W3C Working Draft http ://www.w3.org/TR/rdf-sparql-query/*.

Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E. & Ciravegna F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Journal of Web Semantics*, **4**.